**NYC : For Hire Vehicles (FHV) – Active**

**Objective:**

To read the NYC FHV dataset using API on scheduled manager and process the data to ingest into S3 location. To perform the SQL analytics, required tables will be created on Athena, which help to perform analytics on data to make informed decisions.

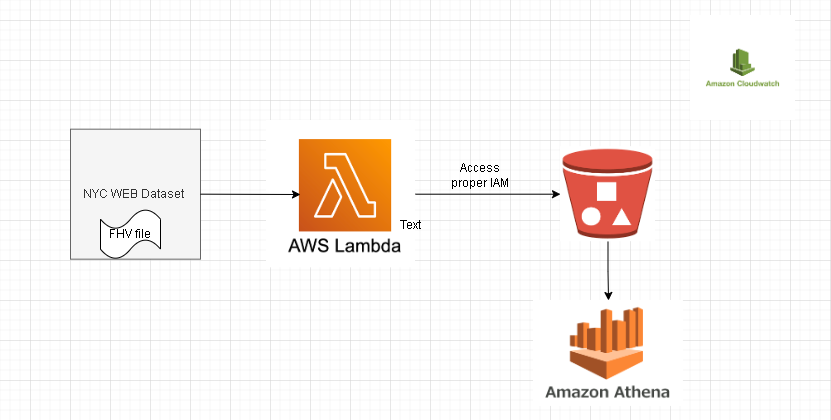
Source Links:

[For Hire Vehicles (FHV) - Active | Socrata API Foundry](https://dev.socrata.com/foundry/data.cityofnewyork.us/8wbx-tsch)

<https://data.cityofnewyork.us/resource/8wbx-tsch.csv>

[For Hire Vehicles (FHV) - Active | NYC Open Data (cityofnewyork.us)](https://data.cityofnewyork.us/Transportation/For-Hire-Vehicles-FHV-Active/8wbx-tsch)

**Data flow Diagram:**



This flow is planned to use the minimum resources and go with serverless, as we are dealing less amount of data and make it available on S3.

**Preparing IAM Roles + Policies:**

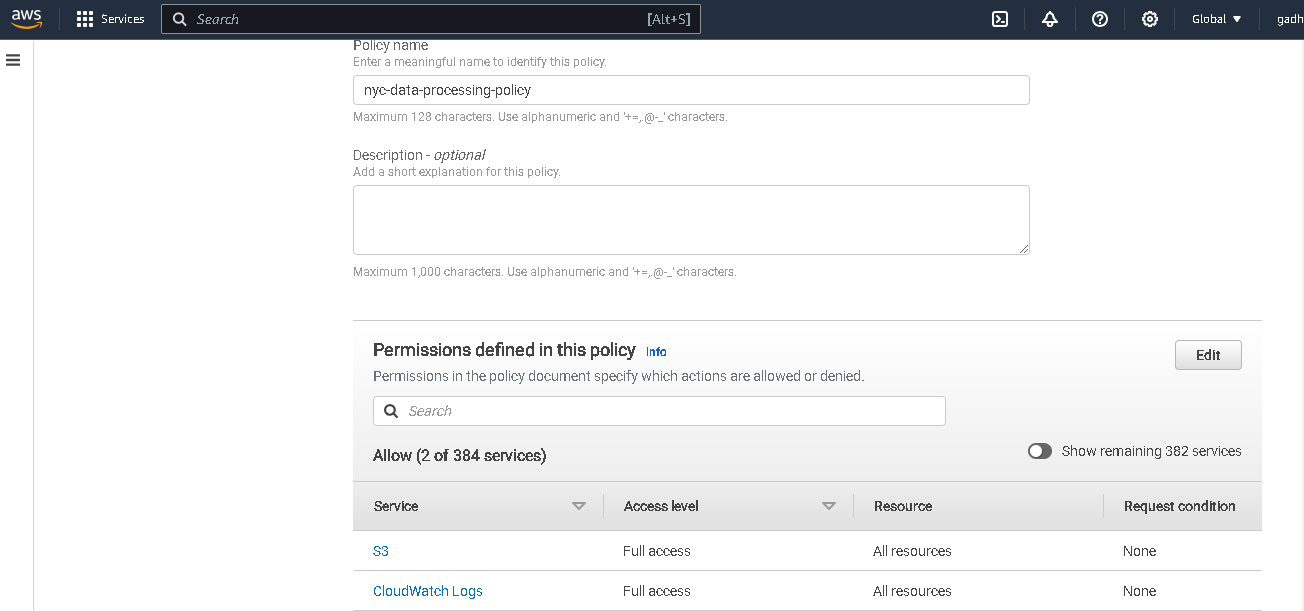
Policy Creation:

For now, selected all polices of the resources (this can be selected only read/write etc.,).

Selected:

S3 permissions to write the files from Lambda

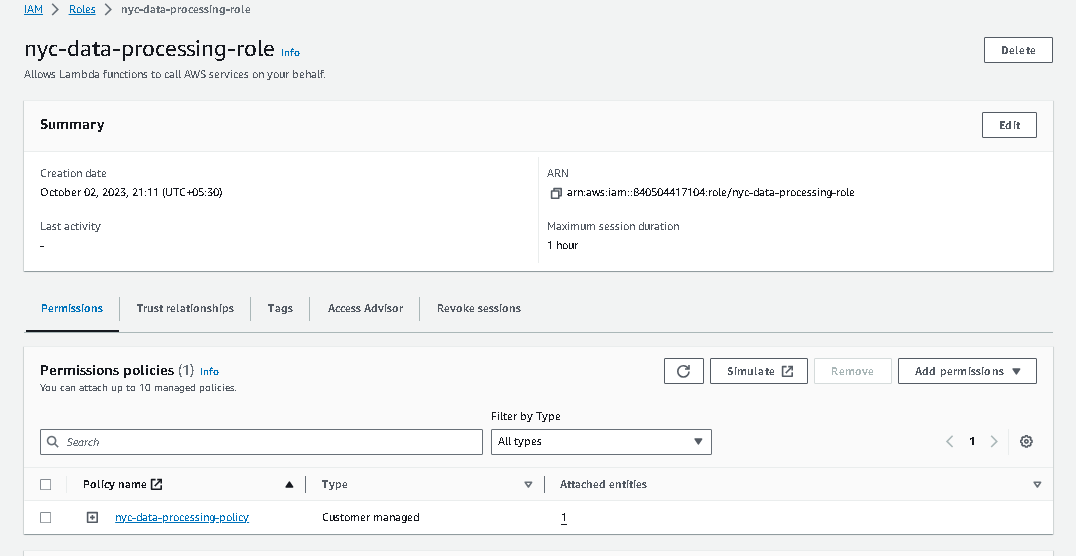
Cloudwatch logs will help to investigate the errors.



Role:

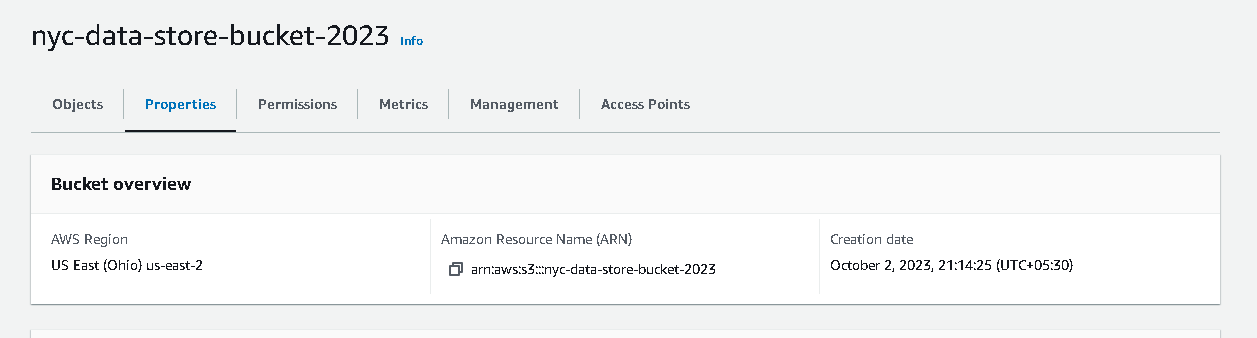
Role created for Lambda service.

Choose the policy created in above screenshot to access S3 and cloudwatch logs.



**S3 Bucket Creation:**

Created S3 bucket to store the data files on incremental fashion. (with default settings)

****

**Lambda Function:**

1.Choose to create the code from Scratch

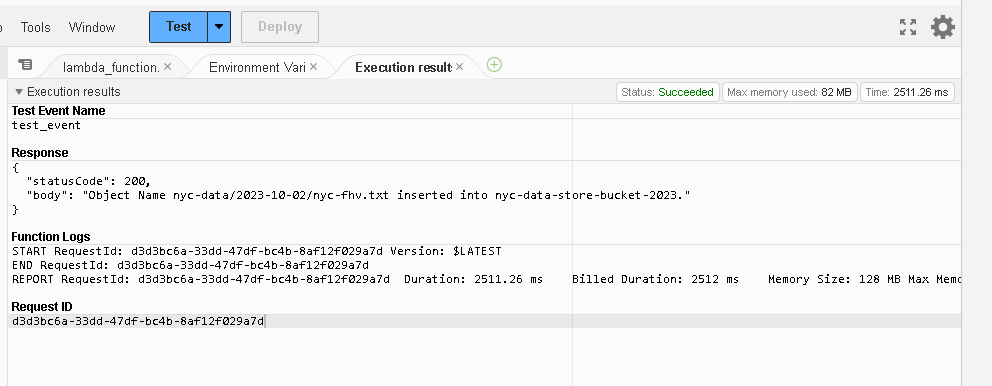
2. opted Python language to write basic function to read the link and post the files into S3 location.

3. Selected permissions created in above screenshots.

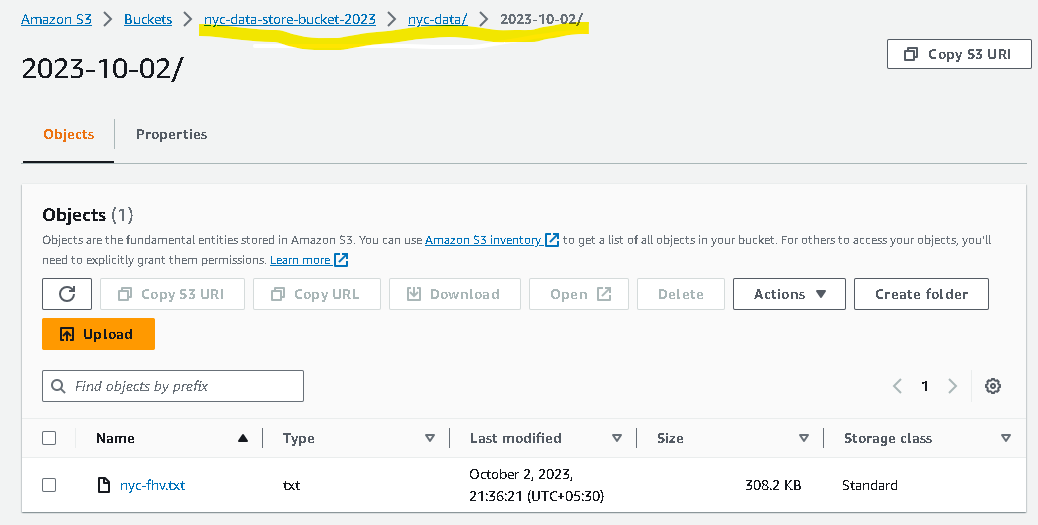
Written below code, which reads the dataset file from Nyc site and copies into S3 as destination.



**Execution result of Lambda Function:**



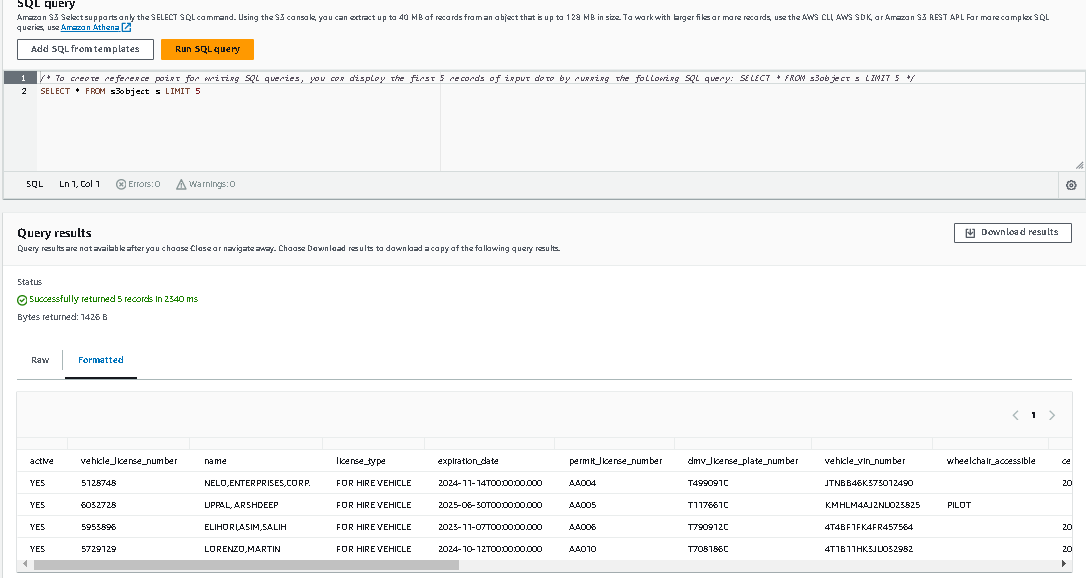
Files Availability on s3 location:



The line underlined will explain that, on every day schedule it will create a date folder and put the files inside that to differentiate the daily file.

In case if we plan to run many times in same day, we can turn the code inside lambda as timestamp.

**Quick Data check using S3 SQL select:**

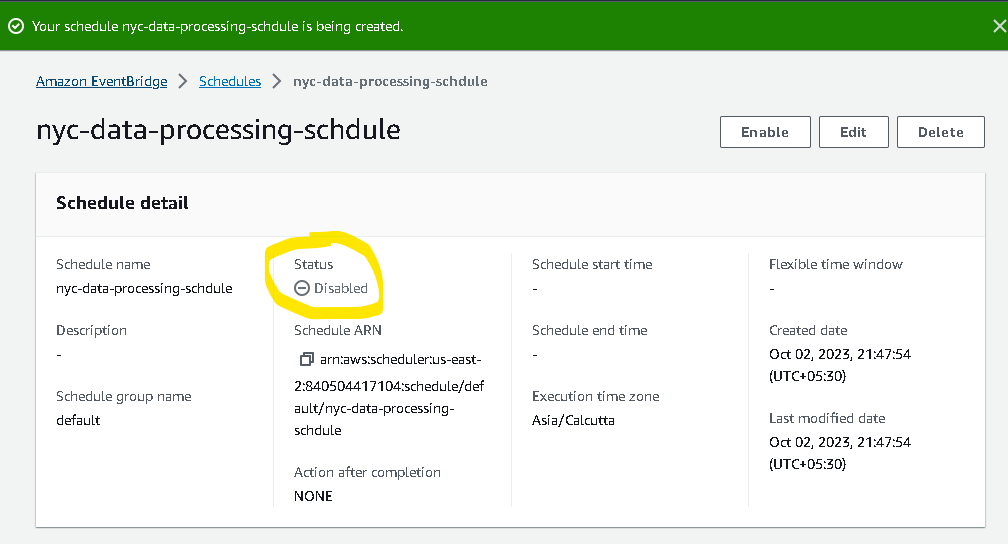


**Scheduling the Lambda Function:**

Under Cloud Watch Service, I have created event by writing Rule.

scheduled Lambda function to run daily at 23:59 minutes

For now, disabled to run and retries also switched-off.

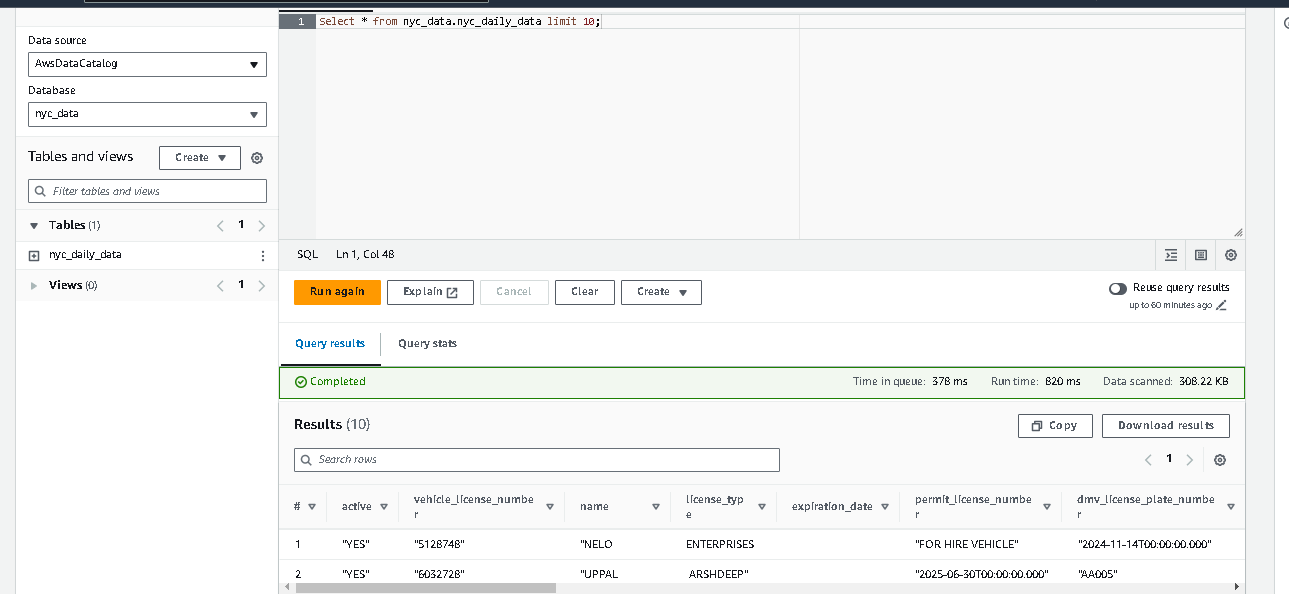


**Athena for SQL analytics:**

Create Table and new database in Athena to read the data from S3 location. Below is the DDL created using UI.



**Quick data check:**



**SQL’s: Written with rough idea**

\*\*\*\*\*\*\*\*\*\*

*select vehicle\_year,VEH, count(\*) as no\_of\_veh\_yr*

*from nyc\_data.nyc\_daily\_data*

*group by 1,2;*

*\*\*\*\*\*\*\*\*\*\**

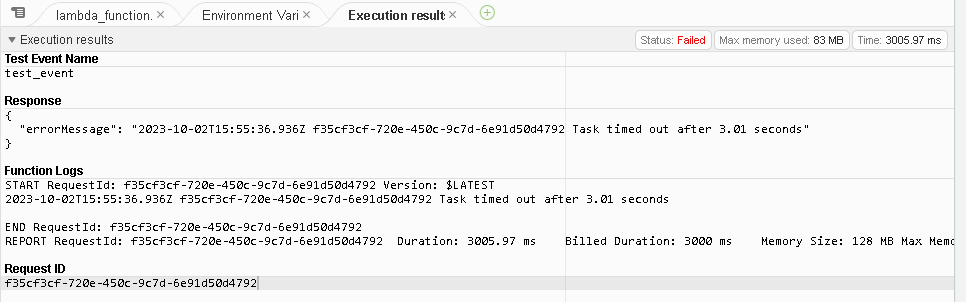
select median(expiration\_Date - certification\_date,0.5)

from nyc\_data.nyc\_daily\_data

group by vehicle\_year;

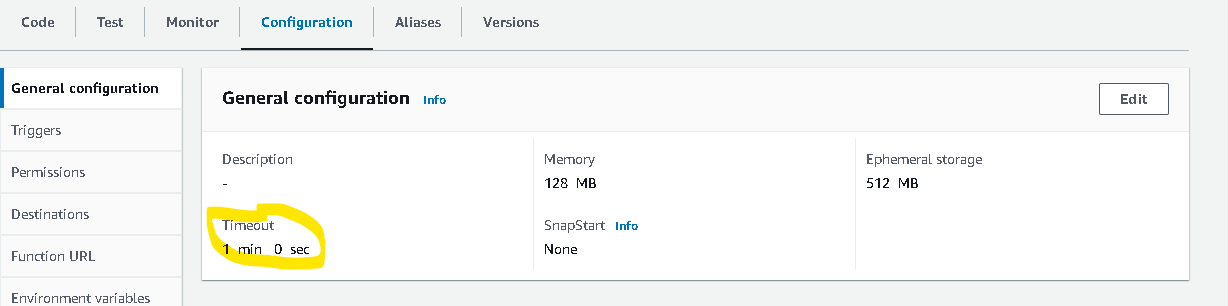
\*\*\*\*\*\*\*\*\*\*\*\*\*

**Issued Faced:**

****

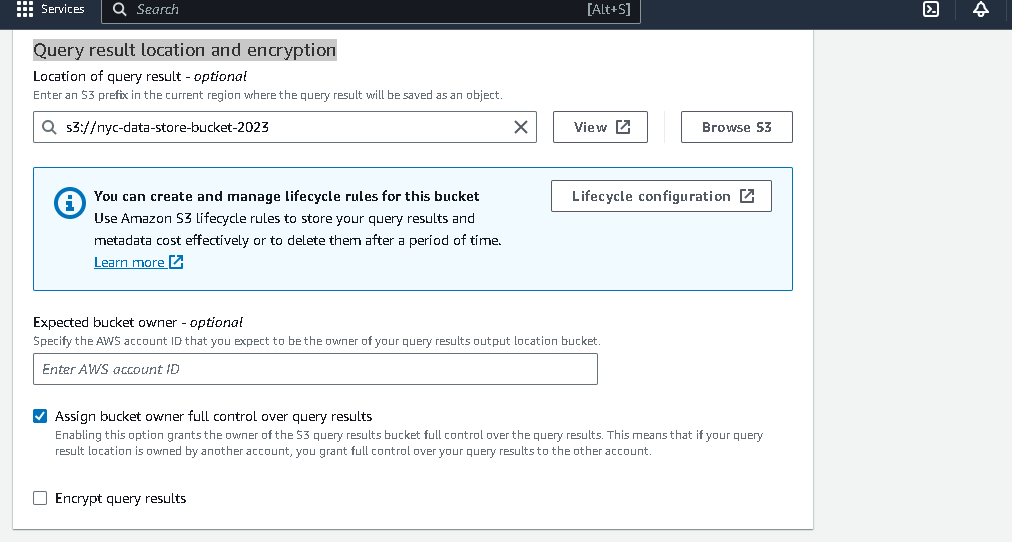
**<Resolution> :** It’s resolved after changing the timeout range to one min from 3 sec which is fault.

Inside Configuration window.



**Issue 2:**

No output location provided. An output location is required either through the Workgroup result configuration setting or as an API input.



**Explain your ratonale for your approach to this task?**

As I mentioned in data flow diagram, I have used minimal resources with serverless (instead of taking provisioning the infra). As NYC data is small Lambda will perfectly fit in the requirement to process the data. In case of huge data, and some transformations to be applied, **AWS glue** could have helped here in better way for ETL operations.

**What else would you do if you had more time?**

1. Granular permissions would have been designed for IAM roles, based on the compliance.
2. S3 life cycle policy for an object– would been implemented to make sure the files are moved to different storage classes to do the cost optimization.

S3: Standard / IA / Glacier

1. If the data was having any PII or PCI information, probably following custom managed keys for encryption will provide more security.
2. Best practice to save the python code into S3 bucket and upload from that location into Lambda. As it helps in case if anyone else needs same code. Can be reutilized.

**Credentials:**

gadhiyan101@gmail.com

Nirman@123456